Comparing French PP-attachment to English, German and Swedish

Martin Volk and Frida Tidström
Stockholm University
Department of Linguistics
106 91 Stockholm, Sweden
volk@ling.su.se

Abstract

The correct attachment of prepositional phrases (PPs) is a central disambiguation problem when parsing natural languages. This paper compares the baseline situation for French as exemplified in the Le Monde treebank with earlier findings for English, German and Swedish.

We perform uniform treebank queries and show that the noun attachment rate for French prepositions is strongly influenced by the preposition *de* which is by far the most frequent preposition and has a strong tendency for noun attachment. We therefore also compute the noun attachment rate for the other prepositions separately as well as for the many complex prepositions that are explicitly marked in this treebank.

1 Introduction

Any computer system for natural language processing has to struggle with the problem of ambiguities. If the system is meant to extract precise information from a text, the ambiguities must be resolved. One of the most frequent ambiguities arises from the attachment of prepositional phrases (PPs). Simply stated, a PP that follows a noun (in French as in English, German or Swedish) can be attached to the preceding noun or to the verb of the same clause.

In the last decade various methods for the resolution of PP attachment ambiguities have been proposed. The seminal paper by Hindle and Rooth (1993) started a sequence of studies for English. Volk (2001; 2002) has investigated similar methods for German. Recently other languages such as Dutch (Vandeghinste, 2002), Swedish (Kokkinakis, 2000; Aasa, 2004), and French (Gaussier and Cancedda, 2001; Gala and Lafourcade, 2005) have followed.

Volk (2006) investigated the attachment tendencies of prepositions in English, German and Swedish. He found that English had the highest overall noun attachment rate followed by Swedish and German. He also showed that the high rate in English was highly influenced by the preposition *of*. From this study he derived a list of criteria for profiling data sets for PP attachment experiments. In the current paper we have applied this list of criteria to French. We have obtained a French treebank and converted it into TIGER-XML so that we can use the same approach as for the other treebanks investigated earlier.

In the PP attachment research for other languages there is often a comparison of the disambiguation accuracy with the results for English. But are the results really comparable across languages? Volk (2006) showed that disambiguation efforts start from very different baselines in English, German and Swedish. In this paper we investigate how French fits into this picture.

2 Background

In their pioneering work Hindle and Rooth (1993) did not have access to a large treebank. Therefore they proposed an unsupervised method for resolving PP attachment ambiguities. A year later Ratnaparkhi et al. (1994) published a supervised approach to the PP attachment problem. They had extracted quadruples V-N-P-N[1] (plus the accompanying attachment decision) from both an IBM computer manuals treebank (about 9000 tuples)

[1] The V-N-P-N quadruples also contain the head noun of the NP within the PP.

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 191–198

and from the Wall Street Journal (WSJ) section of the Penn treebank (about 24,000 tuples). The latter tuple set has been reused by subsequent research, so let us focus on this one.[2] Ratnaparkhi et al. (1994) used 20,801 tuples for training and 3097 tuples for evaluation. They reported on 81.6% correct attachments.

But have they solved the same problem as (Hindle and Rooth, 1993)? What was the initial bias towards noun attachment in their data? It turns out that their training set (the 20,801 tuples) contains only 52% noun attachments, while their test set (the 3097 tuples) contains 59% noun attachments. The difference in noun attachments between these two sets is striking, but Ratnaparkhi et al. (1994) do not discuss this (and we also do not have an explanation for this). But it makes obvious that Ratnaparkhi et al. (1994) were tackling a problem different from Hindle and Rooth (1993) given the fact that their baseline was at 59% guessing noun attachment (rather than 67% in the Hindle and Rooth experiments).

Of course, the baseline is not a direct indicator of the difficulty of the disambiguation task. We may construct (artificial) cases with low baselines and a simple distribution of PP attachment tendencies. For example, we may construct the case that a language has 100 different prepositions, where 50 prepositions always introduce noun attachments, and the other 50 prepositions always require verb attachments. If we also assume that both groups occur with the same frequency, we have a 50% baseline but still a trivial disambiguation task.

In reality the baseline puts the disambiguation result into perspective. If, for instance, the baseline is 60% and the disambiguation result is 80% correct attachments, then we will claim that our disambiguation procedure is useful. Whereas if we have a baseline of 80% and the disambiguation result is 75%, then the procedure can be discarded.

So what are the baselines reported for other languages? And is it possible to use the same extraction mechanisms for V-N-P-N tuples in order to come to comparable baselines across languages?

For English, Volk (2006) had used sections 0 to 12 of the WSJ part of the Penn Treebank (Marcus et al., 1993) with a total of 24,618 sentences for his experiments. He computed a noun attachment rate of 75% over all common nouns (see section 5 for a definition of the noun attachment rate, NAR). This is a surprisingly high number. One reason for this high baseline stems from the fact that he queried for all sequences noun+PP as possibly ambiguous whereas previous research looked only at such sequences within verb phrases. Since he has done the same for all other languages, this is still worthwhile.

For German he had mainly used the large NEGRA and TIGER treebanks with a total of 60,000 trees. He computed a 60% noun attachment rate for common nouns over these treebanks. And for Swedish he had looked at a part of the Talbanken treebank with 6100 trees, which also resulted in a rate of 60% for regular nouns (while he computed significantly higher values for deadjectival nouns (69.5%), and deverbal nouns (77%). Taken together this results in a NAR of 64%.

Now we want to compare these results with a French treebank. We have obtained the French newspaper treebank Le Monde developed at Université Paris 7. The development and the major annotation decisions are described in (Abeillé et al., 2003). The treebank is accompanied by guidelines detailing the annotation decisions concerning the morpho-syntactic annotation (Abeillé and Clément, 2003), the constituent structure annotation (Abeillé et al., 2004) and the functional labels.

The Le Monde treebank consists of two parts. Part one contains 20,500 trees with constituent structure nodes but no functional information. In contrast, part two does contain functional information added to 9,300 trees. The treebank is distributed in a proprietary XML format. We have converted the treebank into TIGER-XML for use with the query tool TIGER-Search.

TIGER-Search is a powerful treebank query tool developed at the University of Stuttgart (König and Lezius, 2002). Its query language allows for feature-value descriptions of syntax graphs. It is similar in expressiveness to tgrep (Rohde, 2005) but it comes with graphical output and highlighting of the syntax trees plus nice frequency tables.

3 Conversion of the Le Monde Treebank to TIGER-XML

TIGER-XML allows the declaration of all annotation features (and the sets of possible values) which will be checked during the import of

[2] The Ratnaparkhi training and test sets were later distributed together with a development set of 4039 V-N-P-N tuples.

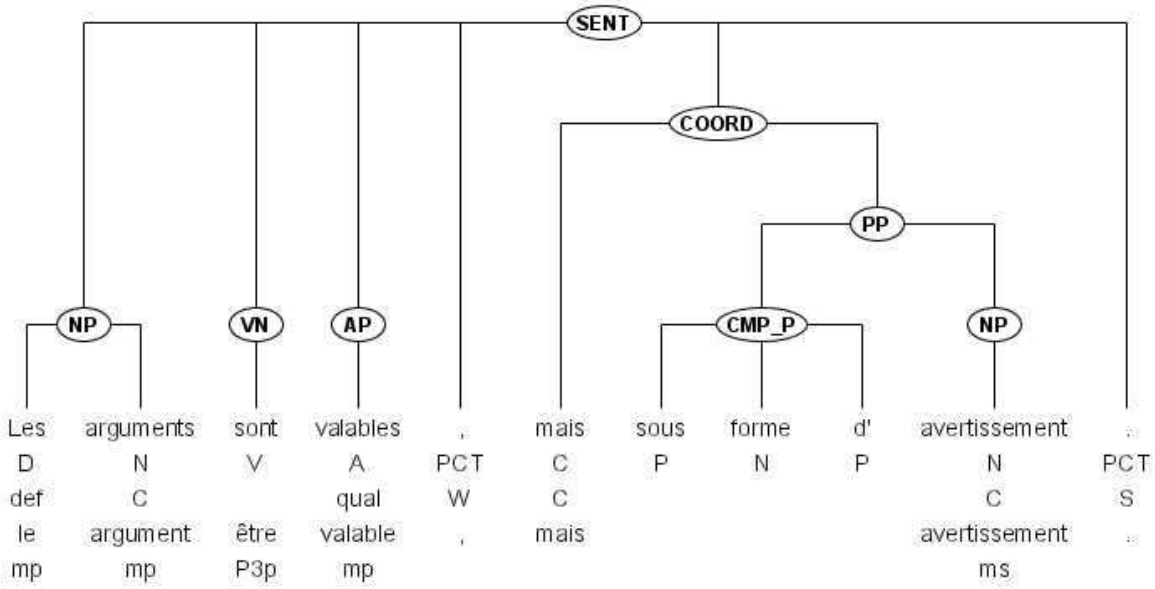

Figure 1: Tree from the Le Monde Treebank with coordination and a compound preposition (CMP_P)

the treebank into the TIGER-Search query tool. Therefore we first collected all word level features and all syntactic features from the French treebank files and compared them against the treebank documentation.

3.1 Word level features

The Le Monde treebank comes with the following word level information: Part-of-Speech tags (main classes and subclasses), morphology information and lemmas.

The developers have made some interesting tokenization decisions. Contracted word forms lead to the insertion of empty tokens (e.g. the preposition *du* gets the lemma *de* and leads to the insertion of an empty token with the lemma *le*). This is an elegant solution to explicitly represent the determiner.[3] Apostrophe contractions are split into two tokens (e.g. *d'un, l'on, c'est*). But in order to capture multi-token units, compounds are specially marked. For example *jusqu'au* is first split into two tokens and then annotated as compound preposition. (Other compounds are annotated accordingly: e.g. *premier ministre* as compound noun and *au contraire* as compound adverb.)

It was particularly difficult to preserve the compound information in the conversion, since TIGER-XML does not provide a representation level between tokens and syntactic nodes. We have therefore decided to use category nodes for grouping compounds. We have introduced the special node label CMP for such compounds (compare to the compounded preposition CMP_P in figure 1).

During the conversion it became clear that the treebank authors have not performed domain checks for the values of the various linguistic features. The treebank contains some undefined and undocumented feature values. For example, we found 13 documented PoS tags: the usual A, N, V, P tags, two pronoun classes (clitic and others), plus a tag each for adverbs, conjunctions, determiners, foreign words, interjections, and punctuation symbols. Finally there is a special tag for prefixes which is used for the first part of hyphenated compounds (for example in *vice-présidente* and in *franco-américain*).[4] But we also found two undocumented PoS tags: "PC" which we suspect stands for Préposition-Conjonction, and "X" for which we have no good guess. In addition there were 11 occurrences of erroneous PoS tags (e.g.

[3] Similar to French, German also has a set of contracted prepositions, e.g. German *im* stands for *in dem*. Sometimes a complex lemma is used for marking such contractions: *in-dem*.

[4] These hyphenated compounds are actually split into two tokens but they are not explicitly marked as compounds. This is strange since so many other types of compounds are explicitly marked in the Le Monde treebank.

ADVP, CC, W, PRE) which we have turned into the undefined tag label "–".

Similar problems of out-of-domain labels occurred also for the PoS subclassification tags. The Le Monde treebank uses 18 tags to subclassify the PoS tags. For example the general pronouns are subclassified into demonstrative, interrogative, possessive, and relative pronouns. In addition to erroneous labels, the SubPoS tags have the unfortunate complication that two of them have double meanings: For example "C" stands for both "Common noun" in combination with nouns (PoS tag = N), and it stands for "Coordination" in combination with conjunctions (PoS tag = C).

The Le Monde treebank comes with complex morphology tags (person, gender, number, tense). We represent them as complex features in TIGER-XML (e.g. "S3p" stands for "3rd person plural, present tense, subjunctive"). This means that one needs to use a regular expression search over the complex features when looking for a specific atomic feature (like person or tense).

TIGER-XML allows us to associate features with non-terminals (i.e. nodes) in the tree. In fact a node label (like NP or CMP_P) is just a feature like any other. Since compounds are represented as non-terminals, we have added the Part-of-Speech subclass, the morphology and the lemma as features. For example, the compound noun *banques centrales* comes with the additional information that it is a common noun (subclass) in feminine plural (morphology) and has the lemma *banque centrale*. Compound prepositions are not divided into subclasses and have no associated morphology information, but they do have a lemma which sometimes differs from the surface form, as for *au profit du* which has the lemma *à le profit de*.

3.2 Syntax level features

The Le Monde treebank comes with 13 documented node labels (e.g. NP, PP, SENT). Most of the constituent grouping follows the traditional strategies. Two deviations are noteworthy: First, there is no VP label for finite verb phrases but rather a label for the verb nucleus. This helps to avoid crossing branches in cases where the subject is located between the verb and the objects. However, infinitive and participle verb phrases are marked.[5]

[5]The avoidance of finite VPs is similar to the annotation in the German TIGER treebank.

preposition	freq	percentage
de, d', des, du	39188	53.2%
à, au, aux	10683	14.5%
en	4779	6.5%
dans	3569	4.8%
par	3091	4.1%
sur	2675	3.6%
pour	2508	3.4%
avec	1573	2.1%
entre	733	1.0%

Table 1: The most frequent French prepositions

Second, the annotation of coordinated structures is strange in that the first conjunct is superior to the second which is introduced with a node labeled COORD (e.g. NP[Christian Blanc COORD[and NP[Eric Frey]]]). The coordination in figure 1 is an example of sentence coordination.

As mentioned above, part two of the treebank contains additional functional information for subject, modifier, and different types of objects. Only constituents with these functions get a functional label. All others are left empty. For example, there is no explicit head information. In this paper we focus on the larger part of the treebank (i.e. the part that lacks the functional information), and we will refer to this first part as the Le Monde treebank hereafter.

4 Prepositions in the Le Monde Treebank

In the Le Monde treebank, there are 73,650 atomic preposition tokens directly dominated by a PP (rather than as part of a compound preposition or some other constituent). They account for 46 preposition types (counted via their lemmas, i.e. *d', des, du* and *de* count as the same preposition)[6]. We present the 9 most frequent prepositions in table 1. These comprise 93.2% of the atomic preposition tokens. As we can see, the preposition *de* strongly dominates the list.

The 46 preposition types are a relatively low number compared to, for instance, German which usually counts around 100 atomic preposition types (Volk, 2001). But a comparison with other French preposition lists confirms this number. For

[6]Note that the query `[cat="PP"] > [pos="P"]` leads to 64 different preposition types. But manual inspection shows that 18 of them are spelling errors (e.g. *ee* instead of *en*), mathematical symbols (+-/) or compound prepositions.

example the French PrepLex Database[7], which contains the merged information from a number of sources (including the syntactic part of PrepNet (Saint-Dizier, 2006)), lists 49 "simple" prepositions (in contrast to multi-word prepositions). 40 of them also occur as prepositions in the Le Monde treebank. The nine remaining ones either do not occur at all in the Le Monde treebank (*circa, confer, versus*) or they occur only with other PoS labels (*dixit, passé, sitôt, touchant, vu, ès*). On the other hand, there are six simple prepositions in the Le Monde treebank which are not listed in PrepLex (*autour, courant, environ, plein, plus, près*). Two of them occur only once as preposition but many times with other PoS labels (*autour* is adverb in 79 cases, and *plein* is adjective 28 times) which leaves some doubt about their status as preposition. The two sentence contexts do not force this interpretation either in our judgement.

In addition to prepositions dominated directly by PPs, the Le Monde treebank contains 21,570 atomic preposition tokens dominated by other categories. Table 2 lists the most frequent categories that dominate prepositions. Column 2 gives the frequency for how often the category contains a preposition, and column 3 gives the percentage relative to the sum of all frequencies in column 2 (including some rare categories not listed in the table).

Not surprisingly compound prepositions lead the list, but also infinitive and participle verb phrases are frequently introduced by a preposition. For example: *un moyen simple VPinf[de prouver cette intention]* (a simple way to prove this intention). Furthermore there are different compounds that contain prepositions: compound adverbs (*à tout prix, aujourd'hui*) , compound nouns (*arrêts de travail, sac à main*), compound verbs (*être en train, rappeler à l'ordre*), and even compounded conjunctions (*pour que, à mesure que*).

There are 5266 compound prepositions in the French treebank. These range from two-token compounds (e.g. *près de*) to seven-token compounds (*d' un bout à l' autre de*). The two-token compounds are mostly combinations with *de* or *à* on the second position. Also *d'ici, d'après, d'abord* with the preposition on the first position are regarded as compound prepositions in this treebank. The three-token compounds are mostly

[7]The French PrepLex can be found at http://loriatal.loria.fr/Resources/PrepLex.txt

category	P freq	P percentage
compound preps	6896	32.0%
infinitive VPs	6817	31.6%
compound adverbs	3868	17.9%
compound nouns	2324	10.8%
participle VPs	583	2.7%
NPs	502	2.3%
compound conj.	202	0.9%
compound verbs	125	0.6%

Table 2: Categories with prepositions

frozen prepositional phrases like *par rapport à, en raison de, à partir de*. The same is true for four-token and longer compounds with the restriction that almost all of them have *de* on the final position *à la fin de, dans le cadre de, au – sein de, de l' autre côté de*. In total there are 460 (!) compound preposition types. The most frequent ones are *par rapport à, il y a* and *près de*. This compares to 206 multi-word prepositions in PrepLex.

It might be surprising that *il y a* is listed as a compound preposition since none of its parts is a preposition. But the annotators of the Le Monde Treebank are not alone in this categorization. It is mentioned by (Grevisse, 1993) and also listed in PrepLex.

Compound prepositions function as heads in PPs in the majority of cases (4665 or 88.6%), but - like atomic prepositions - they also introduce infinitive VPs (in 7.8% of the cases; e.g. as in *[de peur de] s'attirer certaines foudres syndicales*) and NPs (in 3.3% of the cases). Such NPs are mostly introduced by *près de, plus de* and *moins de* as for example in *[près de] 1 million de tonnes*.

Unfortunately the treebank authors have not performed rigid consistency checks over the annotation of compound prepositions. For example, the sequence *en début de* as in *en début de semaine* is annotated once as a compound preposition, but 9 times it is annotated as a nested PP *[en NP[début PP[de NP[semaine]]]]*. In one case it is even annotated as a part of a compounded adverb in *en début de matinée*.

Interestingly coordination of PPs is relatively rare. The treebank contains only 27 cases of a preposition which is dominated by the COORD category. About half of them are comparative constructions with the preposition *comme* like in *au Royaume-Uni comme en Allemagne*.

According to (Pedersen et al., 1989), there is a

clear distinction between a PP attribute and a PP verb complement in French. The PP *i klubben* in the Swedish sentence "*Hon deltog aktivt i diskussionerna i klubben*" is ambiguous since it can be attached either to the verb *deltog* or to the noun *diskussionerna*. However, the French preposition *de* is principally used for PP attributes: *Elle participait activement aux discussions du club* (She participated actively in the discussions of the club).

A different preposition would be used for a PP adverbial: *Elle participait activement aux discussions au club* (She participated actively in the discussions in the club). The prepositions *de* (noun attachment) and *à, dans, sur* (verb attachment) accent different aspects in the examples in table 3.

5 Computing Noun Attachment Rates

Now we would like to determine the attachment tendency for the various French prepositions and the overall attachment tendency for French prepositions. We do that by computing the noun attachment rate (NAR) according to the following formula:

$$NAR = \frac{freq(noun + PP, noun\_attachm)}{freq(noun + PP)}$$

We assume that all PPs in noun+PP sequences which are not attached to a noun are attached to a verb. This means we ignore the very few cases of such PPs that might be attached to adjectives (as for instance the PP in *tard dans la soirée* (late in the evening)).

We compute the frequencies with TIGERSearch queries over the Le Monde treebank. Our experiments for determining attachment tendencies proceed along the following lines. We first query for all sequences of a noun immediately followed by a PP. With the dot being the precedence operator, we use the query:

```
[pos="N"] . [cat="PP"]
```

This query gives us the frequency of all ambiguously located PPs. We find that 35,787 out of 79,011 PPs (45.3%) in this treebank are in such an ambiguous position. These numbers include both common and proper nouns and PPs with all kinds of prepositions. We disregard the fact that in certain clause positions a PP in such a sequence cannot be verb-attached and is thus not ambiguous. For example, a French noun+PP sequence in subject position is not ambiguous with respect to PP attachment since the PP cannot attach to the verb. Similar restrictions apply to Enlish, German and Swedish.

Since we distinguish common nouns and proper nouns in our investigations, we used a refined version of the above query which includes the SubPoS value with either "C" or "P".

In order to determine how many of these sequences are annotated as noun attachments, we query for noun phrases that contain both a common noun and an immediately following PP. This query looks like:

```
#np:[cat="NP"] >  #pp:[cat="PP"] &
#np >* #n:[pos="N" & subpos="C"] &
#n . #pp
```

All strings starting with # are variables and the $>$ symbol is the dominance operator. So, this query says: Search for an NP (and call it #np) that directly dominates a PP, and the NP also dominates (directly or indirectly) a noun which is immediately followed by the PP.

6 NAR Results for French

The first query finds that there are 34,476 occurrences of a PP immediately following a common noun in the Le Monde treebank. The second query results in 28,294 cases of a noun phrase dominating both the PP and the noun. In addition we find 395 cases of a (higher) PP which dominates the noun and the (lower) PP. So we add these two numbers (395 + 28,294) and divide by the number of all occurrences.

This leads to a NAR for common nouns followed by atomic prepositions of 83.2%, which is very high. French clearly has a tendency to attach the PP to the preceding noun. One reason must be that French produces genitive-like structures, compounds and measures with the help of the preposition *de*. Let us have a closer look at the attachment tendencies of the different prepositions in table 4.

Column 1 lists the lemmas of the 9 most frequent French prepositions. Column 2 contains the frequency of the preposition being the head of a PP in a (common) noun attachment context (like in query 2), while column 3 contains the frequency (being the head of a PP) in an ambiguous position (like in query 1). The rightmost column lists the NAR for each preposition (i.e. the ratio of the two previous columns given as percentage).

Clearly the high NAR for French is mainly due to the preposition *de* which accounts for more

	Verb attachment	Noun attachment
(1)	*Elle a construisit un hôpital à Nice* She constructed a hospital in Nice	*Elle mourut dans un hôpital de Nice* She died in a hospital in Nice
(2)	*Il avait fait une tache sur le mur* He had made a spot on the wall	*Il se rappela la tache du mur* He remembered the spot on the wall
(3)	*Cela a causé un scandale dans les années trente* This caused a scandal in the thirties	*Le livre décrit un scandale des années trente* The book describes a scandal of the thirties

Table 3: Examples of noun vs verb attachments (taken from (Pedersen et al., 1989))

prep.	freq P N-att	freq P	NAR
de	23726	24387	97.3%
entre	202	282	71.6%
sur	537	930	57.7%
avec	190	375	50.7%
à	1308	2668	49.0%
par	289	618	46.7%
pour	341	768	44.4%
en	656	1502	43.7%
dans	291	855	34.0%

Table 4: The NAR for the most frequent French prepositions (relative to common nouns)

than half of the preposition tokens in the treebank. Only two more of the frequent prepositions show a clear noun attachment tendency: *entre* and *sur*. If we omit *de* from the computation of the NAR, we end up with a balanced situation, i.e. a NAR of 50% for all the remaining prepositions. We should also mention that there are some rarer prepositions with high NARs (an example is *sans* which occurs 75 times as preposition in the treebank and has a NAR of 73.3%).

This situation is very similar to English where the preposition *of* has a noun attachment rate of 99% and is very frequent. If *of* is omitted from the calculation, English in fact shows a tendency towards verb attachment.

Furthermore we find that the NAR for proper nouns followed by atomic prepositions is 42% in the French treebank. This is also in line with Volk's (2006) findings in the other languages. Proper nouns don't take prepositional complements and attributes as often as common nouns. For example, for German he found a NAR of around 20% for proper nouns.

If we look at French compound prepositions, the picture changes. We find a NAR of only 37.7% for compound prepositions which follow common nouns in the Le Monde treebank. We don't have any comparative figures for English, German and Swedish since compound prepositions are not marked in the treebanks for these languages.

7 Conclusions

Our findings put other results for French PP attachment resolution into perspective. If our NAR of 83.2% is a fair assessment of the attachment tendency of French PPs in ambiguous positions, then any lower accuracy scores based on automatic disambiguation are meaningless. A simple program can achieve 83.2% correct attachments by always predicting noun attachment for all ambiguously located PPs (that are headed by atomic prepositions).

Consider for example (Gaussier and Cancedda, 2001) who have tested their disambiguation method "against 900 manually annotated sequences of nuclei from the newspaper Le Monde". Since they give no reference to the Le Monde treebank, we assume that they used different annotated material, accidentally from the same newspaper. They report on results of 73.5% correct PP attachments. But they have only looked at V N P sequences which is different from our approach. Unfortunately they do not give the NAR for their data.

Our results are also interesting for general linguistic insights into the behavior of French prepositions. Second language learners could profit as well. The profiling of the prepositions with respect to their attachment tendencies tells a lot about the usage options.

In future work we would like to test the methods proposed in (Volk, 2001; Volk, 2002) for the resolution of German PP attachment ambiguities against the French treebank. Furthermore we would like to make contrastive studies on prepositions based on parallel treebanks. This will lead

to an increased understanding of cross-language prepositional correspondences and help in building machine translation systems.

8 Acknowledgements

We would like to thank Anne Abeillé for making the French Le Monde treebank available to us. Part of this research was done while the first author was a visiting researcher at Macquarie University in Sydney. We gratefully acknowledge financial support through the Australian HCSNet.

References

Jörgen Aasa. 2004. Unsupervised resolution of PP attachment ambiguities in Swedish. Master's thesis, Stockholm University. Combined C/D level thesis.

Anne Abeillé and Lionel Clément. 2003. Annotation morpho-syntaxique. Les mots simples - les mots composés. Corpus Le Monde. Technical report, LLF, UFRL, Paris 7.

Anne Abeillé, Lionel Clément, and Francois Toussenel. 2003. Building a Treebank for French. In Anne Abeillé, editor, *Building and Using Parsed Corpora*, volume 20 of *Text, Speech and Language Technology*, chapter 10, pages 165–187. Kluwer, Dordrecht.

Anne Abeillé, François Toussenel, and Martine Chéradame. 2004. Corpus Le Monde. Annotations en constituants. Guide pour les correcteurs. Technical report, LLF, UFRL, Paris 7.

Nuria Gala and Mathieu Lafourcade. 2005. Combining corpus-based pattern distributions with lexical signatures for PP attachment ambiguity resolution. In *Proc. of SNLP-05, 6th Symposium on Natural Language Processing*, Chiang Rai, Thailand.

Eric Gaussier and Nicola Cancedda. 2001. Probabilistic models for PP-attachment resolution and NP analysis. In *Proc. of ACL-2001 CoNLL-2001 Workshop*, Toulouse. ACL.

Maurice Grevisse. 1993. *Le Bon Usage (Refondu par André Goose)*. Editions Duculot, Paris.

D. Hindle and M. Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Dimitrios Kokkinakis. 2000. Supervised PP-attachment disambiguation for Swedish. *Nordic Journal of Linguistics, Special Issue on Cognitive Approaches to Language*, 23(2):191–213.

Esther König and Wolfgang Lezius. 2002. The TIGER language - a description language for syntax graphs. Part 1: User's guidelines. Technical report.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

John Pedersen, Ebbe Spang-Hanssen, and Carl Vikner. 1989. *Fransk Universitetsgrammatik*. Esselte Studium, Akademiförlaget. Translated by Olof Eriksson and Lars Lindvall.

A. Ratnaparkhi, J. Reynar, and S. Roukos. 1994. A maximum entropy model for prepositional phrase attachment. In *Proceedings of the ARPA Workshop on Human Language Technology*, Plainsboro, NJ, March.

Douglas L. T. Rohde, 2005. *TGrep2 User Manual*. MIT. Available from http://tedlab.mit.edu/~dr/Tgrep2/.

Patrick Saint-Dizier. 2006. PrepNet: a Multilingual Lexical Description of Prepositions. In *Language Resources and Evaluation Conference (LREC)*, pages 877–885, Genova. European Language Resources Association (ELRA).

Vincent Vandeghinste. 2002. Resolving PP attachment ambiguities using the WWW (abstract). In *Computational Linguistics in the Netherlands*, Groningen.

Martin Volk. 2001. *The automatic resolution of prepositional phrase attachment ambiguities in German*. Habilitationsschrift, University of Zurich.

Martin Volk. 2002. Combining unsupervised and supervised methods for PP attachment disambiguation. In *Proc. of COLING-2002*, Taipeh.

Martin Volk. 2006. How bad is the problem of PP-attachment? A comparison of English, German and Swedish. In *Proc. of ACL-SIGSEM Workshop on Prepositions*, Trento, April.